Assignment 2

Stavan Mehta, Saniya Khinvasara, Foram Trivedi

December 27, 2024

1. Regex

Question: Write a Python script using the re module to extract all phone numbers in the format `(XXX) XXX-XXXX` or `XXX-XXX-XXXX` from the given text.

2. Tokenization in SpaCy

Question: Using SpaCy, write a script to tokenize the sentence: `"SpaCy is a powerful NLP library for tokenization."` Print each token and its corresponding whitespace status (is_space).

3. Spacy Pipeline

Question: Create a custom component in a SpaCy pipeline that counts the number of tokens in a document. Integrate this component into the pipeline and process the sentence: `"Customizing a SpaCy pipeline can be very useful."` Print the token count at the end.

4. Part of Speech Tagging in SpaCy

Question: Write a SpaCy script to process the text: `"NLP is fascinating, and SpaCy makes it even more interesting."` Extract and print each word along with its part of speech and dependency relation.

5. Named Entity Recognition in SpaCy

Question: Using SpaCy, write a script to perform Named Entity Recognition (NER) on the following text: `"Elon Musk, the CEO of Tesla, was born on June 28, 1971, in Pretoria, South Africa."` Print each entity, its label, and the description of the label.

6. One Hot Encoding, Bag of Words, TF-IDF

Question: Write a Python program using the sklearn library to demonstrate One Hot Encoding, Bag of Words, and TF-IDF on the following text corpus: `["Data science is amazing", "Machine learning is part of data science"]` Display the resulting vectors for each method.